IMAGE-BASED CHARACTER RETRIEVAL FROM MOVIES

Instructor
PhD. Mai Tiến Dũng

Nguyễn Nguyên Khôi – 21521009
Đỗ Minh Khôi - 21521007

Table of Contents

1. Problem introduction

- **Input/Output**
- **Applications**

Problem introduction

Input:

- A database (index) of a movie
- A set of query images of a character in that movie

Output:

- All shots in which the character appear

Constraint:

- Not a cartoon movie
- Query images must be cut from the movie

Problem introduction

Requirements:

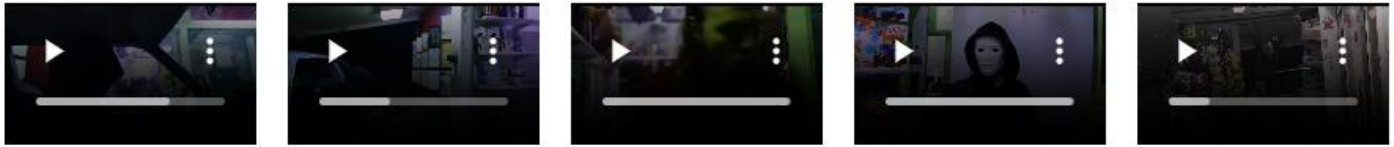
- The result must cover all true shots, including ones in which the person is not clear (faces are unable to detect, too much illumination, etc.)
- As many relevant shots retrieved as possible

Problem introduction

Example: Like Me (2017)

- Input:

- Output:



Problem introduction

Applications:

- Quickly rate a character's performance without watching the entire movie
- Plays a crucial role in key-event summarization task: retrieving important shots from a movie

2. Methods

- **Person detection: YOLOv8**
- **Feature extraction:**
  - **DINO v2**
  - **OpenAI CLIP**
- **Building index for each movie**

Person detection: YOLOv8

YOLOv8-l (Jan 10, 2023 by Ultralytics):

- For person detection in frame
- Input: A single frame (RGB image)
- Output: bounding boxes of persons in frame
- No. of parameters: ~43.7M parameters

Feature extraction: DINOv2

DINOv2 (Maxime et al. 2023):

- A pre-trained visual model with different Vision Transformers architectures
- Learn visual features of images that can be used for various tasks: classification, semantic segmentation, instance retrieval, depth estimation, etc.
- Pre-training curated data: LVD-142M, comprised of ImageNet-22k, ImageNet-1k/train, Google Landmarks, Caltech-101/train, Food-101/train, etc + uncurated data
- Paper: DINOv2: Learning Robust Visual Features without Supervision

Feature extraction: DINOv2

dinov2-small:

- No. of parameters: ~22.1 million
- Embedding dimension: (257, 384), data type: float32
- This embedding is still quite large

  ➔ replace each column with its mean value

  ➔ new embedding dimension: (1, 384)

Feature extraction: OpenCLIP

Contrastive Language-Image Pre-Training (CLIP) (Alec et al. 2021):

- Trained to understand images and text together
- Associate images with their textual descriptions
- Interpret many visual concepts in multiple languages
- Tasks: zero-shot classification, linking images to phrases without re-training
- Applications: image descriptions, image classification, content moderation, image searching, etc.

Feature extraction: OpenCLIP

clip-vit-base-patch16:

- ViT-B/16 Transformer as an image encoder (patch size=16x16)
- Masked self-attention Transformer as a text encoder
- Trained on publicly available image-caption data, crawled from various websites
- Paper: Learning Transferable Visual Models From Natural Language Supervision

Building index for each movie

Facebook AI Similarity Search (Faiss):

- Allows for efficient search of vectorized multimedia documents similar to a given set of vectorized query documents
- Vectors can be compared using L2 (Euclidean) distances or dot products
- Similarity: Lowest L2 or highest normalized dot product (cosine)
- Supports quantizing giant documents for smaller storage and faster search (e.g. HNSW, NSG)

Building index for each movie

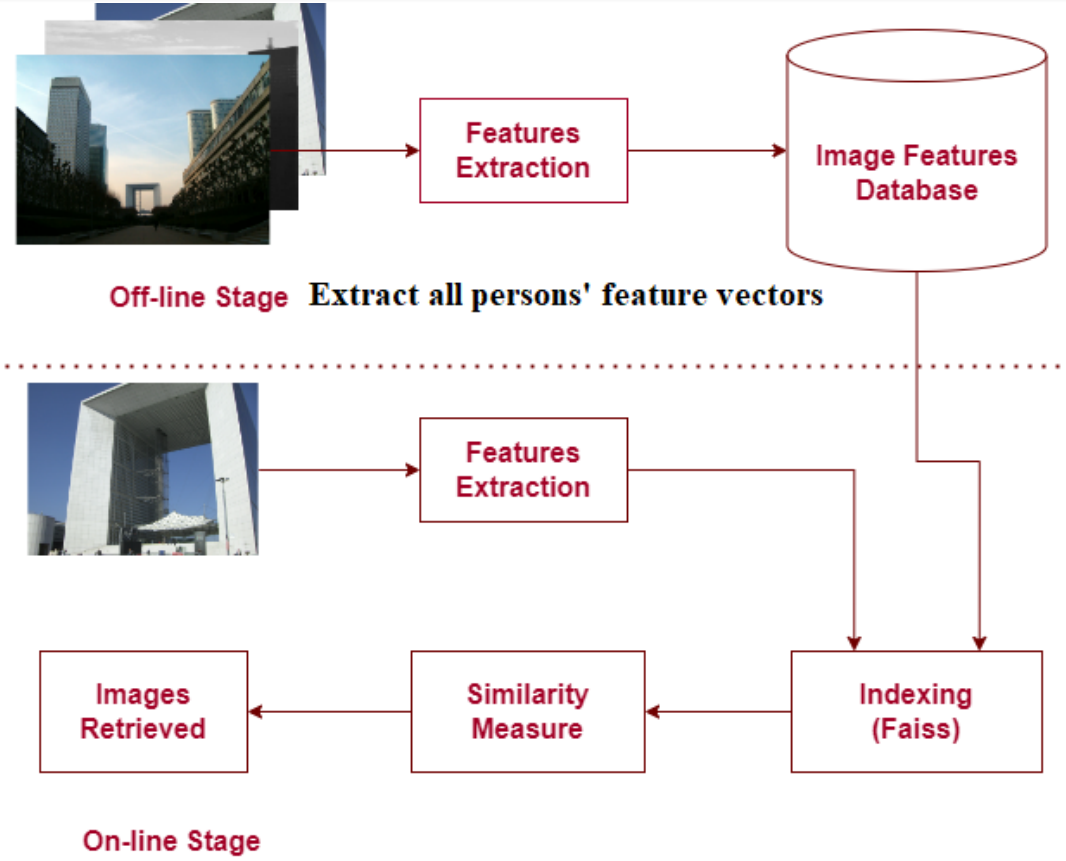
Facebook AI Similarity Search:

- Metric: FlatL2: $L2(u,v) = \sqrt{\sum(ui - vi)^2}$

with u = (u1, u2, u3, .. um), v = (v1, v2, v3,...,vm)

- Index: IndexFlatL2

Building index for each movie



Source: github.com/KhaLee2307/image-retrieval

Building index for each movie

- Persons' info is stored in a .csv file
- Persons in all frames are represented as feature vectors, stored in an array
- Each vector's position in the array is recorded and shown in the "Person" column

Person	Shot	Scene	Full
0	1	1	losing_ground-1-shot_1
1	1	1	losing_ground-1-shot_1
2	1	1	losing_ground-1-shot_1
3	1	1	losing_ground-1-shot_1
4	1	1	losing_ground-1-shot_1
5	1	1	losing_ground-1-shot_1
6	1	1	losing_ground-1-shot_1
7	1	1	losing_ground-1-shot_1
8	1	1	losing_ground-1-shot_1
9	1	1	losing_ground-1-shot_1
10	1	1	losing_ground-1-shot_1
11	1	1	losing_ground-1-shot_1
12	1	1	losing_ground-1-shot_1
13	1	1	losing_ground-1-shot_1
14	1	1	losing_ground-1-shot_1

3. Experiment

- **Dataset: TRECVID_MSUM_2022**
- **Metrics: precision & recall**

Experiment: Dataset

Dataset: TRECVID_MSUM_2022:

Movie	No. scenes	No. shots	No. frames
Like Me	28	870	22440
Memphis	47	259	21650
Losing Ground	40	302	22600
Liberty Kid	56	1048	31803
Calloused Hands	58	965	26774

Experiment: Dataset

Dataset: TRECVID_MSUM_2022:

- frames_5fps:

- All frames ➔ All persons ➔ feature vectors (DINOv2 or CLIP) ➔ saved to an array ➔ saved to an index, all persons stored in a {movie_name}_{db}.csv file

Experiment: Dataset

Dataset: TRECVID_MSUM_2022:

- Query images for Losing Ground:
  - Sara:

- Images ➔ persons ➔ feature vectors (DINOv2 or CLIP) ➔ saved to an array

Experiment: Dataset

Ground truth: example of Sara in Losing Ground:

1	Scene	Shot	Full
2	1	1	losing_ground-1-shot_1
3	1	3	losing_ground-1-shot_3
4	1	5	losing_ground-1-shot_5
5	1	7	losing_ground-1-shot_7
6	1	9	losing_ground-1-shot_9
7	1	10	losing_ground-1-shot_10
8	1	12	losing_ground-1-shot_12
9	1	14	losing_ground-1-shot_14
10	1	16	losing_ground-1-shot_16
11	1	17	losing_ground-1-shot_17
12	1	18	losing_ground-1-shot_18
13	2	1	losing_ground-2-shot_1
14	2	2	losing_ground-2-shot_2
15	2	3	losing_ground-2-shot_3
16	2	4	losing_ground-2-shot_4
17	2	5	losing_ground-2-shot_5
18	2	6	losing_ground-2-shot_6
19	2	7	losing_ground-2-shot_7
20	2	8	losing_ground-2-shot_8

Experiment: Metrics

- Metrics:
  - Precision: $\frac{number\ of\ true\ retrieved\ \textbf{shots}}{number\ of\ \textbf{retrieved}\ \textbf{shots}}$
  - Recall $= \frac{number\ of\ true\ retrieved\ \textbf{shots}}{number\ of\ \textbf{relevant}\ \textbf{shots}}$
  - k = top number of **most relevant persons** in the movie

Experiment: Metrics

Experiment on Kiya and Burt in Like Me:

- Kiya:

- Burt:

Experiment: Metrics

Experiment on Memphis (Willis) and Losing Ground (Sara):

- Memphis:
  - Willis:

- Losing Ground:
  - Sara:

Experiment: Metrics

Experiment on Derrick in Liberty Kid:

- Derrick:

Experiment: Metrics

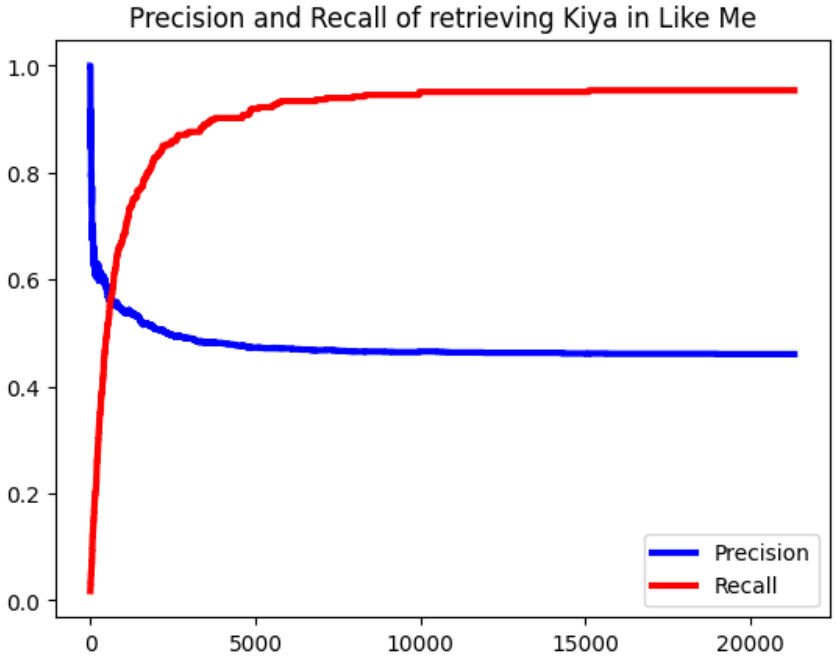
Experiment on Debbie and Byrd in Calloused Hands:

- Debbie:

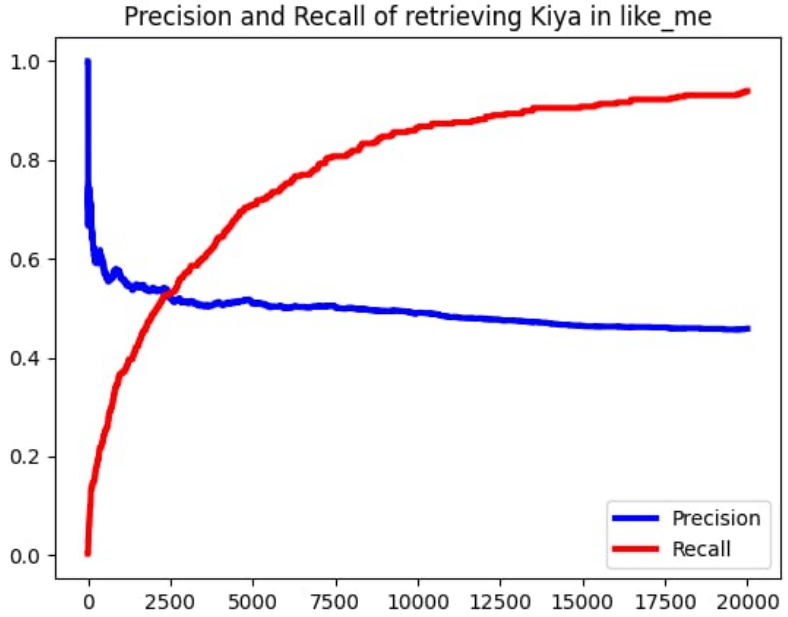
- Byrd:

Experiment: Metrics

DINOv2

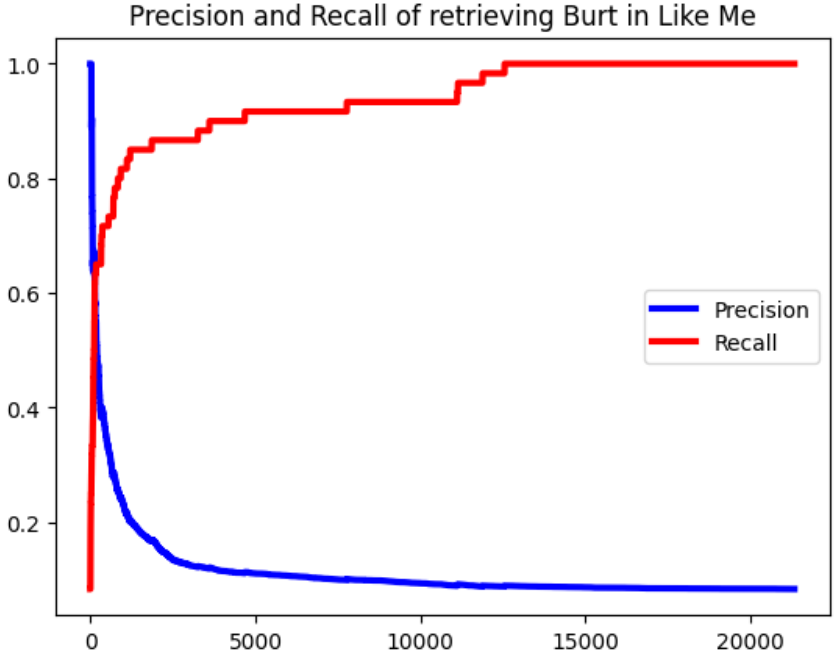


CLIP

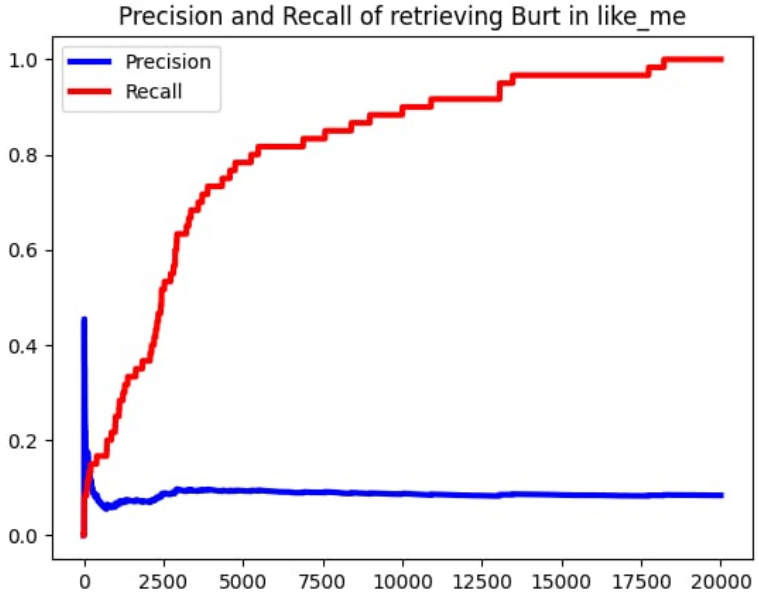


Experiment: Metrics

DINOv2

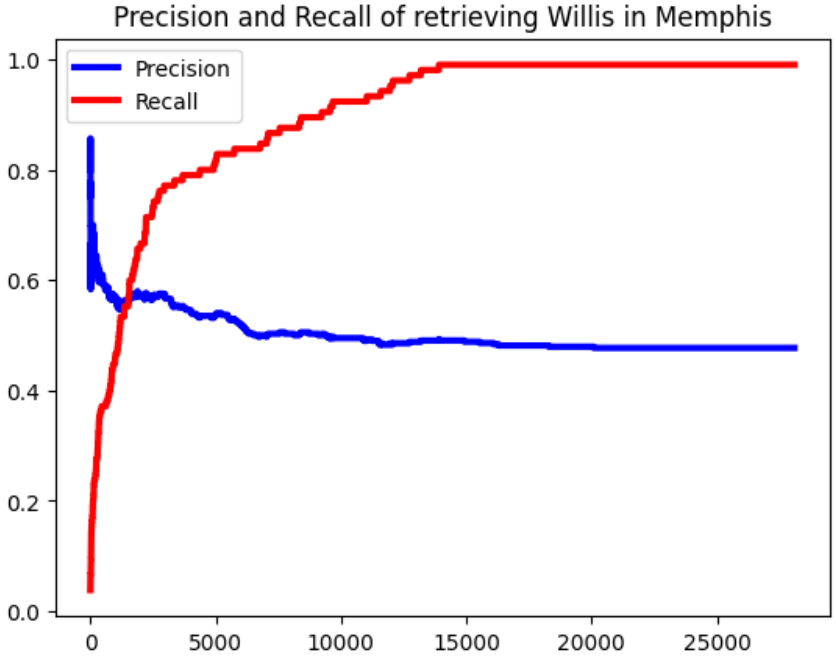


CLIP

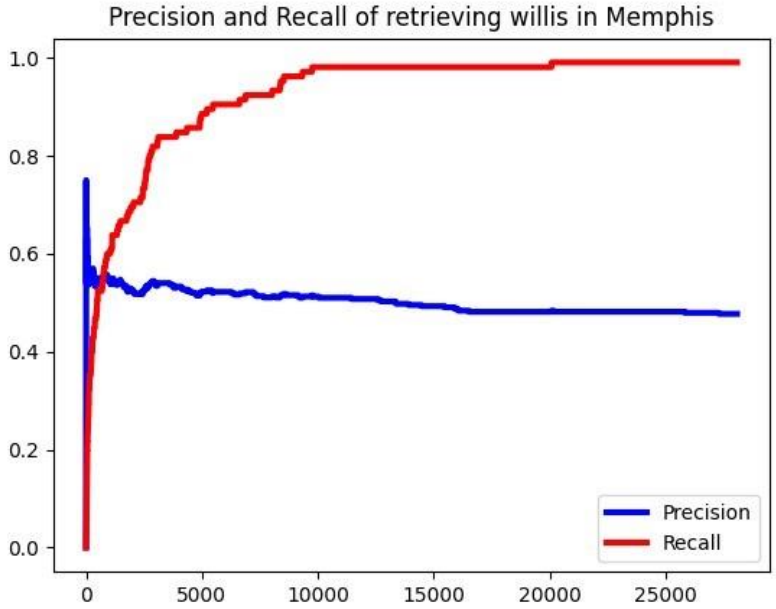


Experiment: Metrics

DINOv2

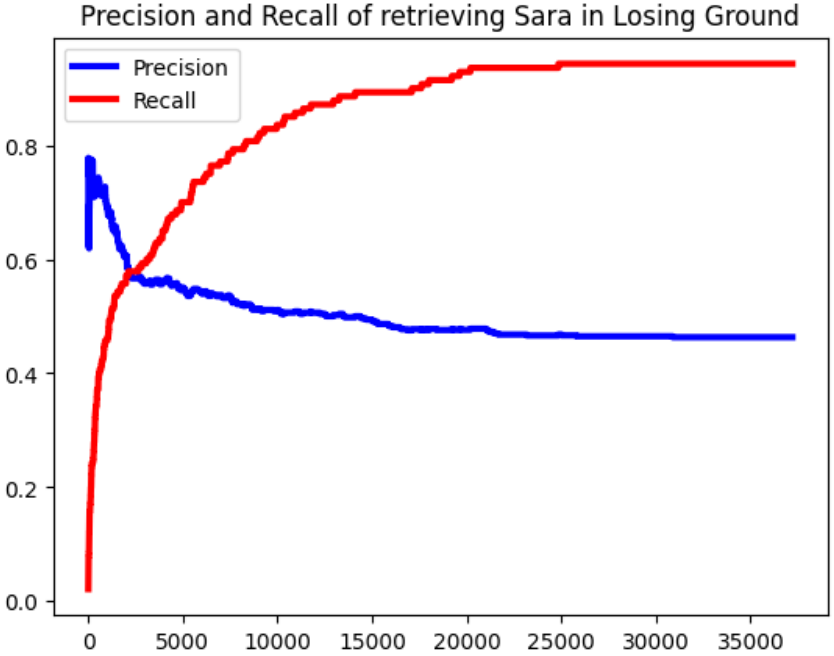


CLIP

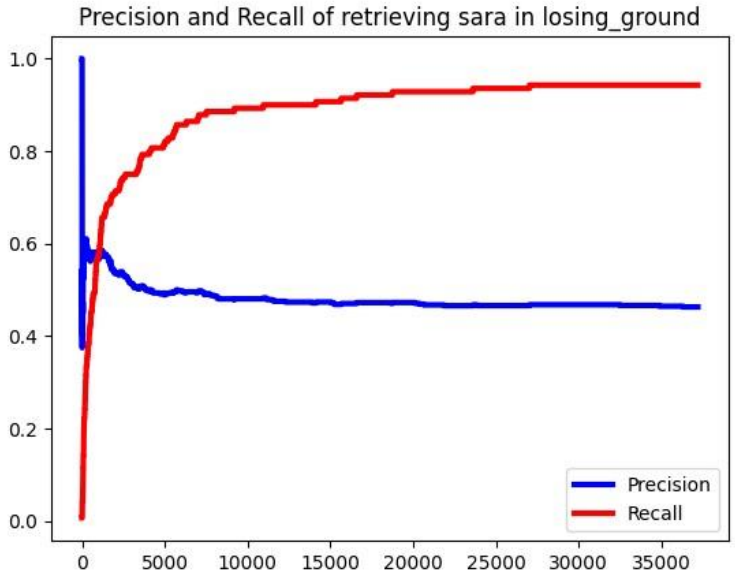


Experiment: Metrics

DINOv2

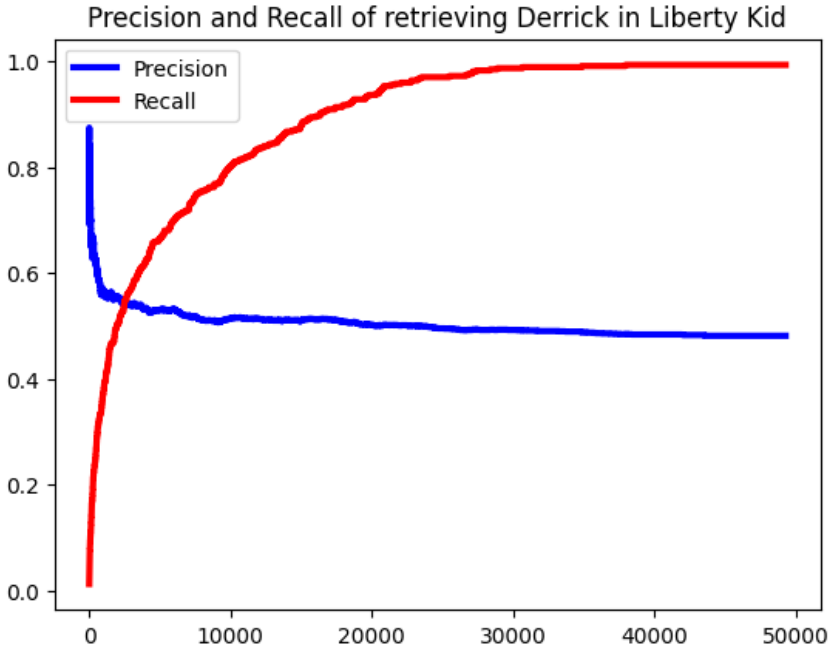


CLIP

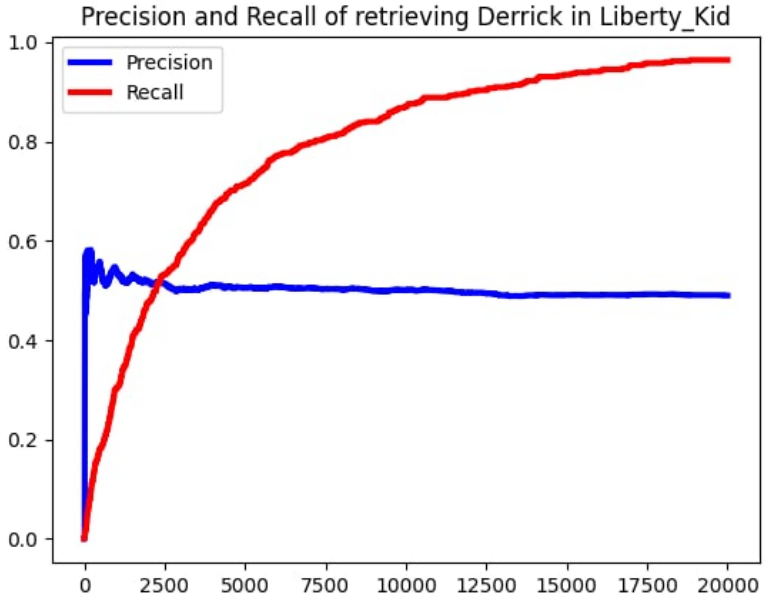


Experiment: Metrics

DINOv2

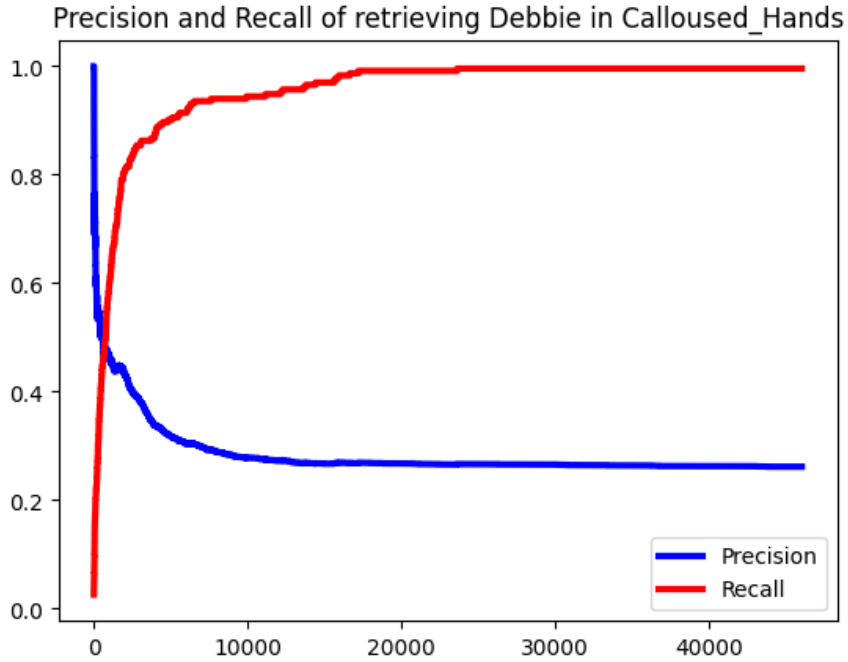


CLIP

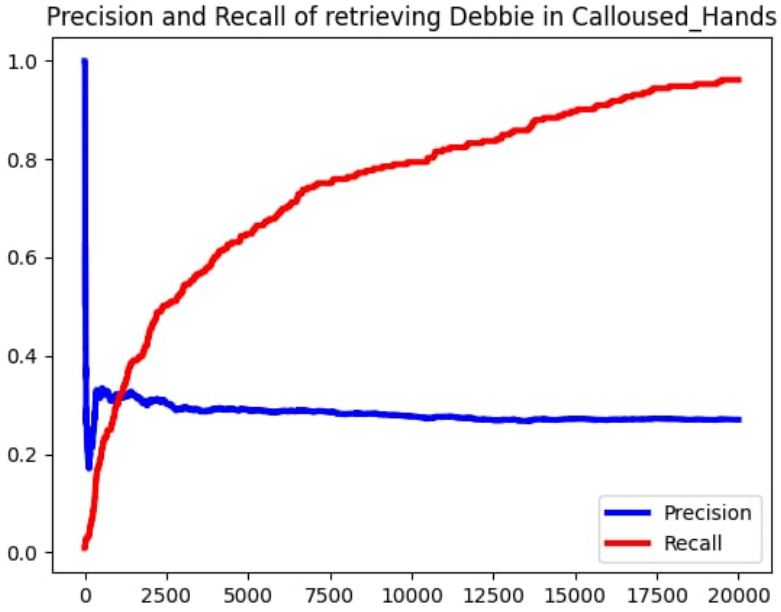


Experiment: Metrics

DINOv2

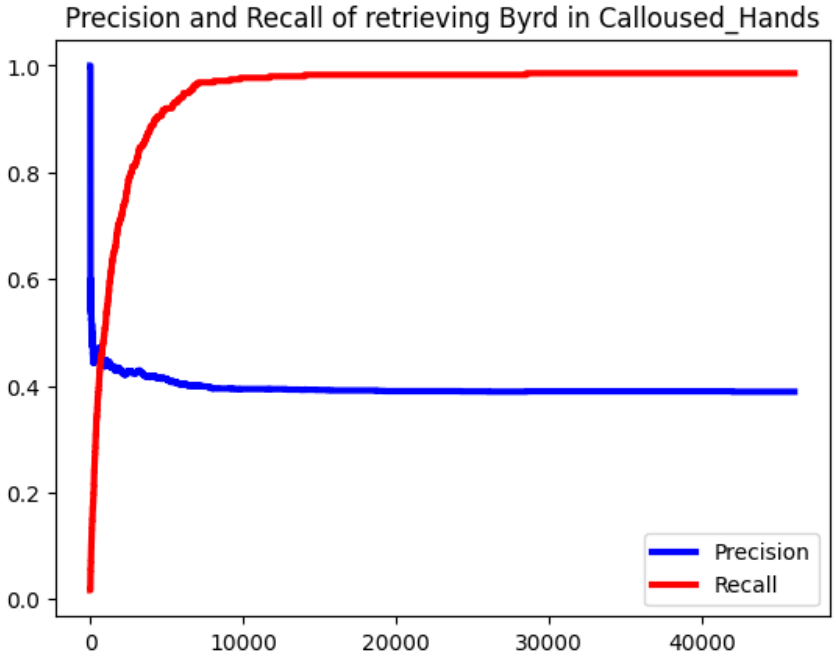


CLIP

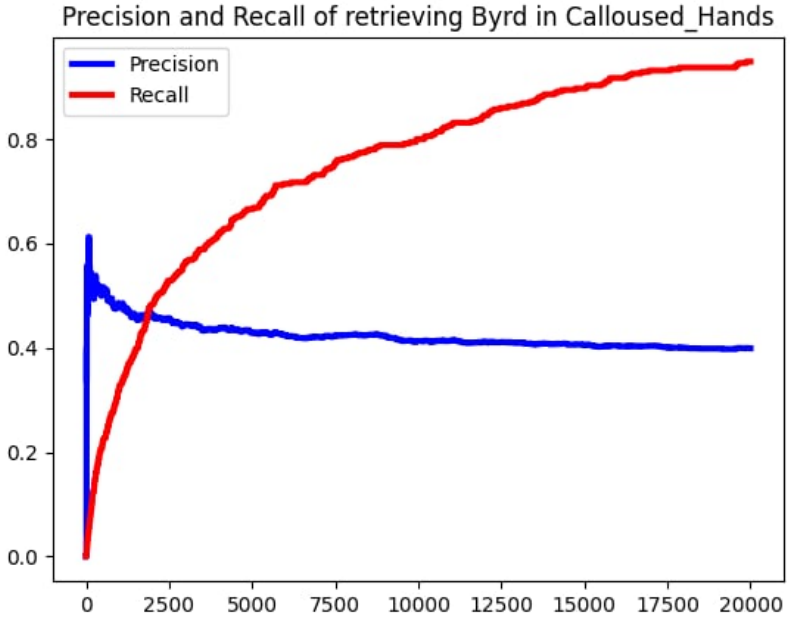


Experiment: Metrics

DINOv2



CLIP



4. Challenges

Challenges

- A lot of frames in which the character does not have a clear appearance (too much illumination, blur, deformed face, masked face, etc.)
- Character's info in the query set may be not enough (lack of images)
- Retrieving shots relevant to a specific query image

➔ Low precision and low recall

5. Demo

References:

[1] Hugging face: dinov2-small

[2] Hugging face: openai-clip-vit-base-patch16

[3] Faiss: A library for efficient similarity search

[4] Image similarity with DINOv2 and Faiss

Q&A